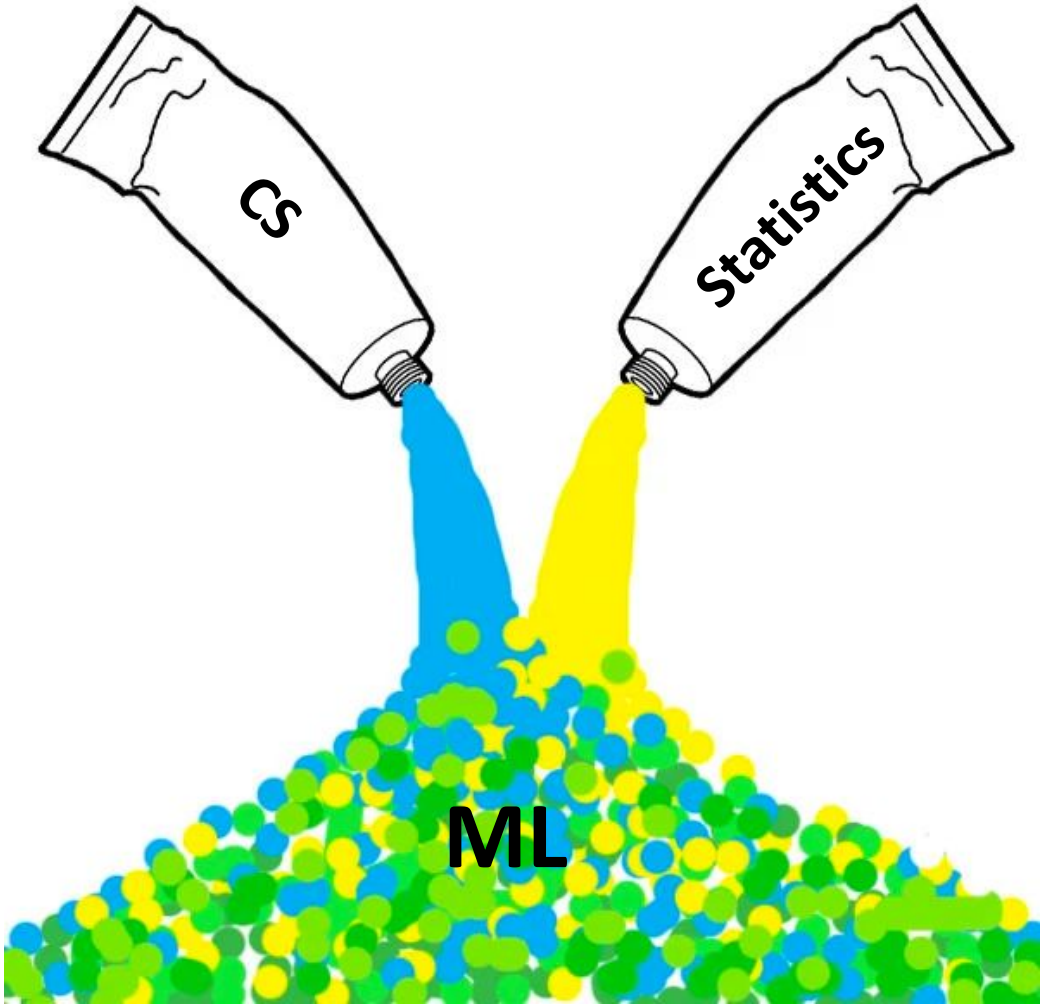
Machine learning? Statistics?

Jeanette Mumford
University of Wisconsin - Madison
Center for Healthy Minds


CS
Statistics
ML

Main ideas

- Is machine learning just fancy statistics?
- Is machine learning better?
- How are people using ML in neuroscience?
- Things to consider when transitioning from voxelwise stats to ML

Main ideas

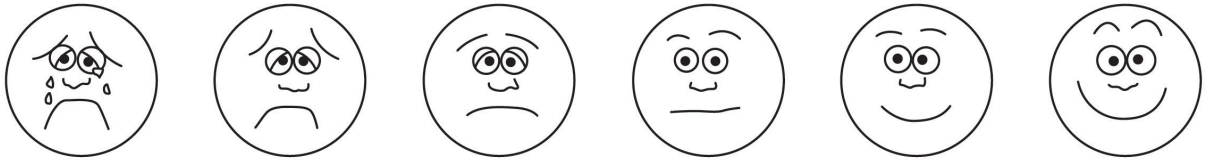
- Is machine learning just fancy statistics?
- Is machine learning better?
- How are people using ML in neuroscience?
- Things to consider when transitioning from voxelwise stats to ML

Simplified comparison

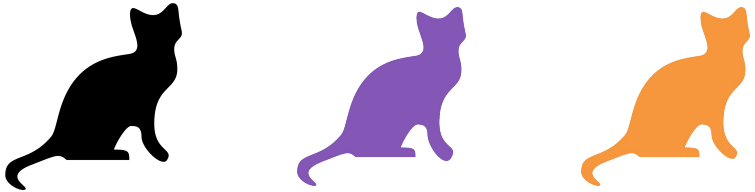
- Statistics
  - Inference
  - Mechanism
- Machine learning
  - Prediction

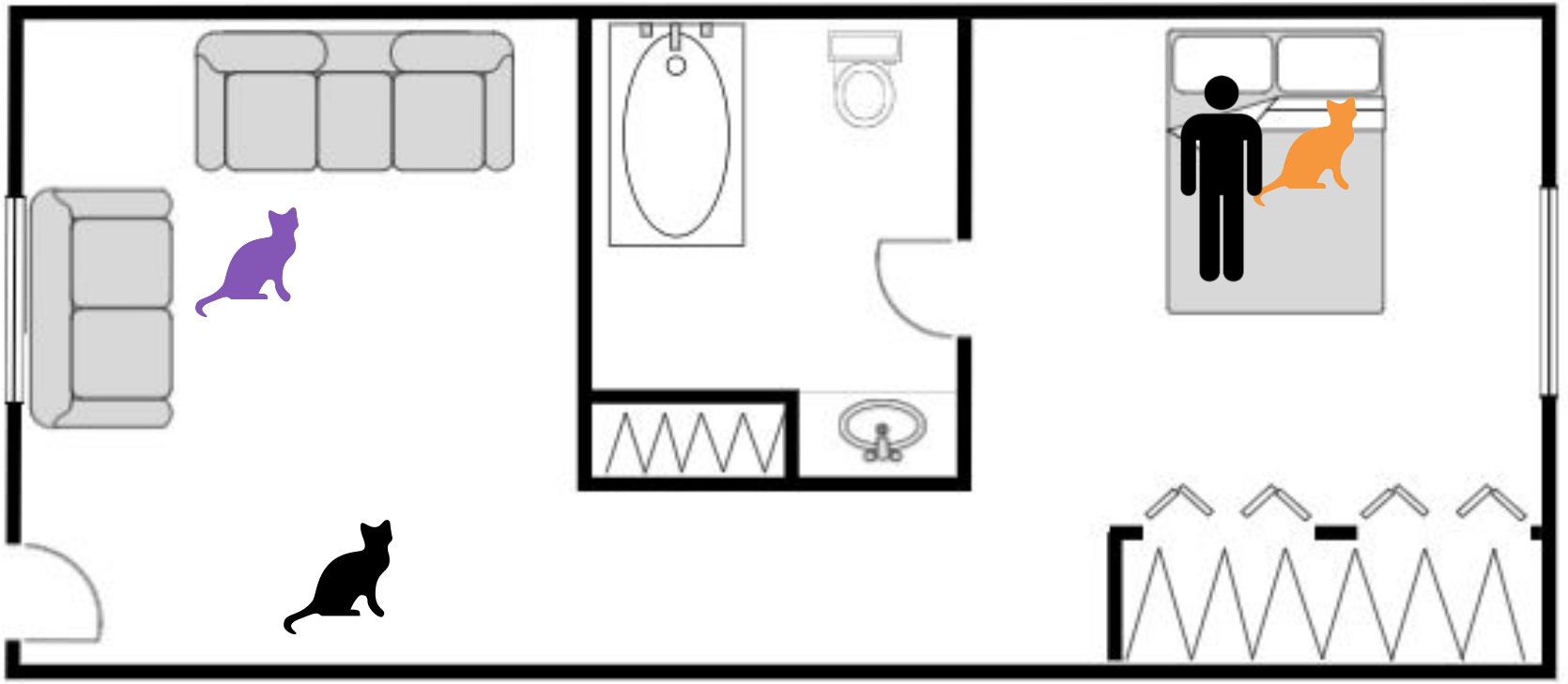
Example

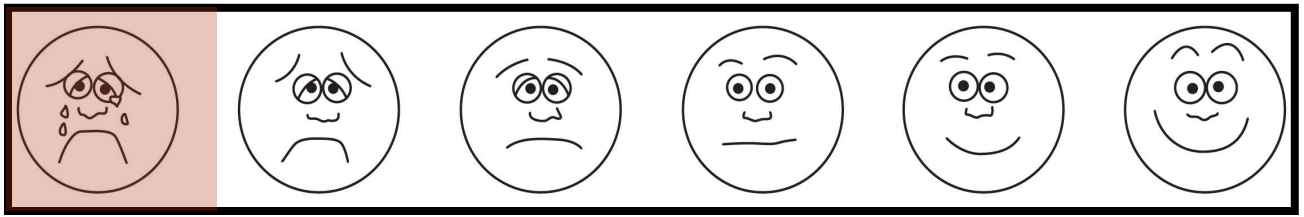
- Measure 1: My current level of happiness over the course of the day

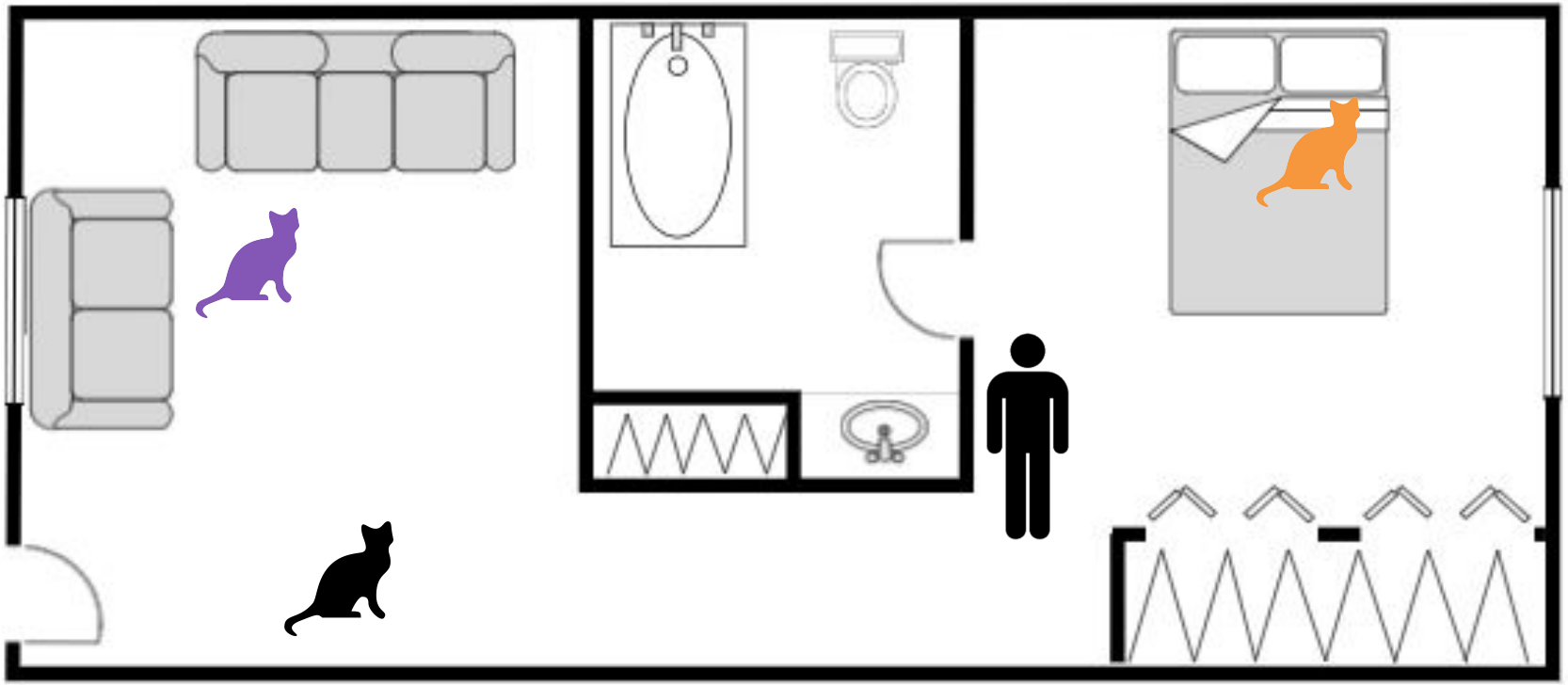


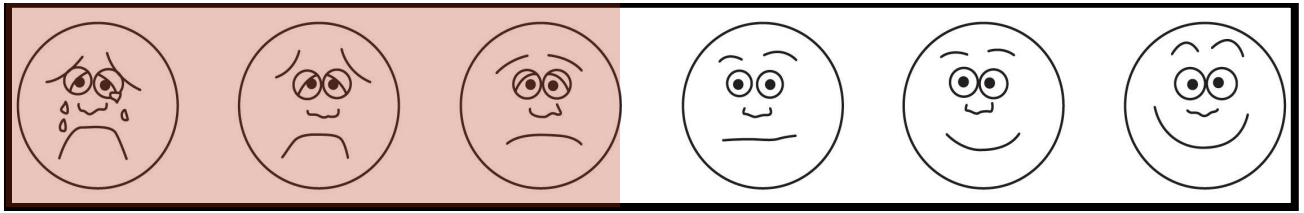
- Measures 2-4: My proximity to each of three different cats

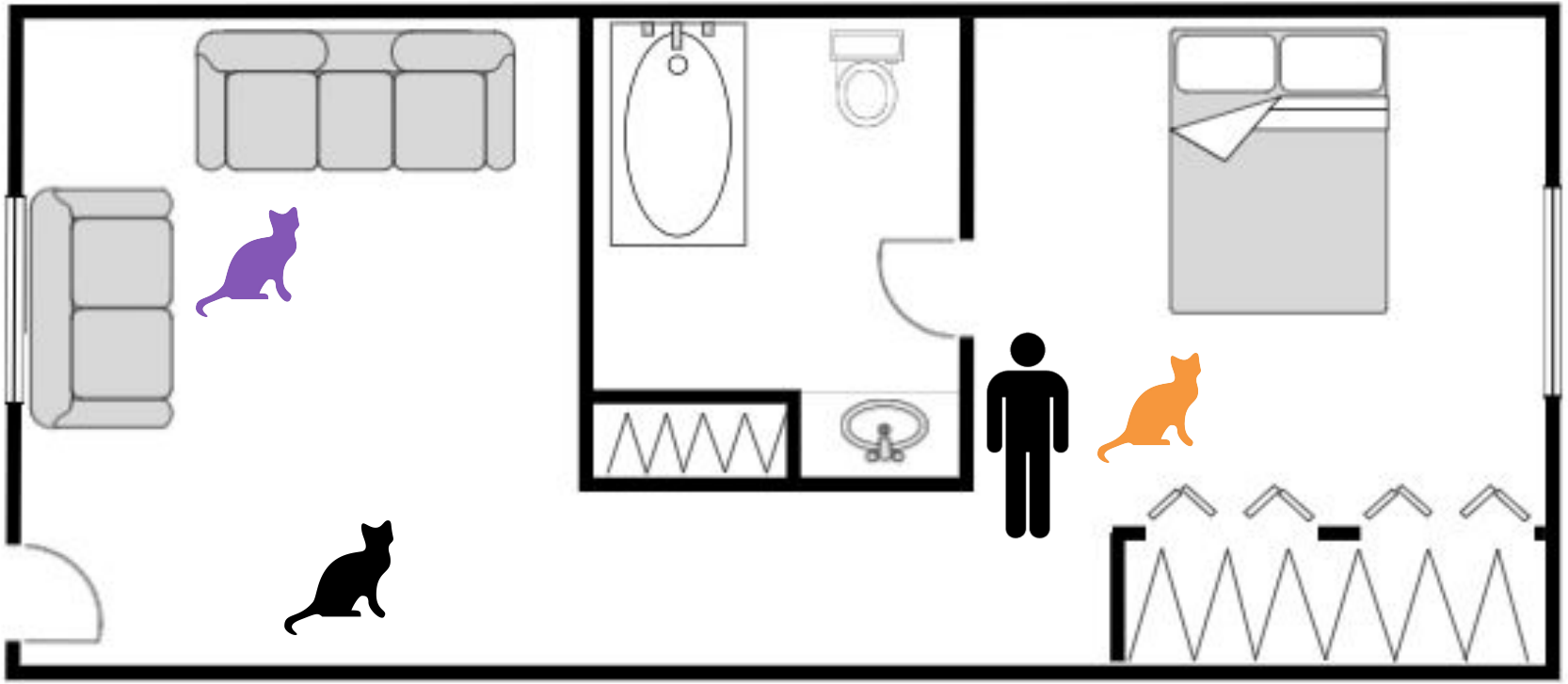


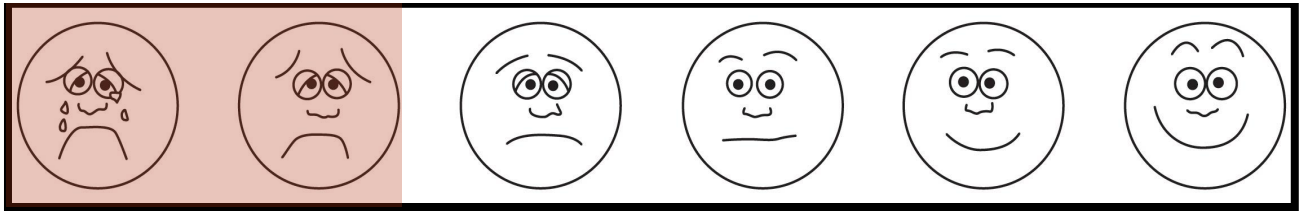


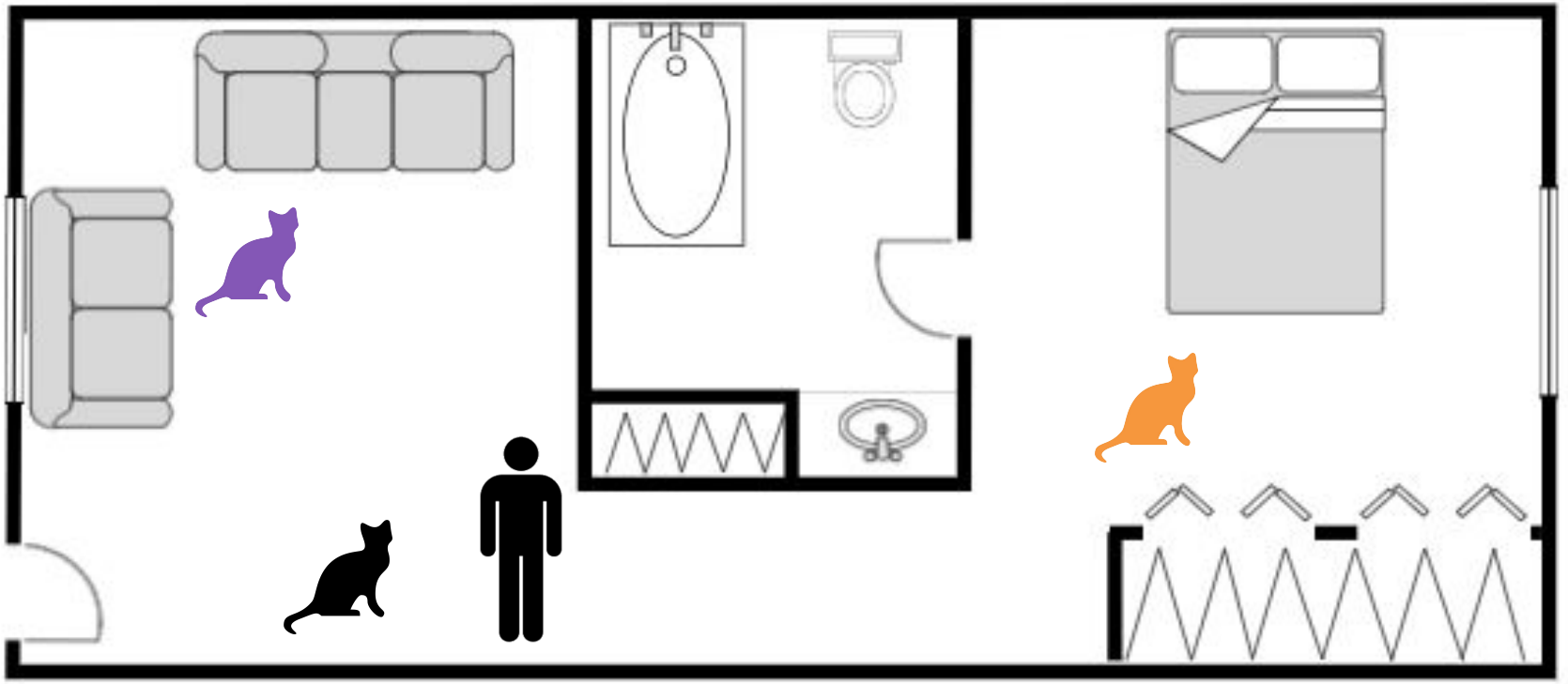


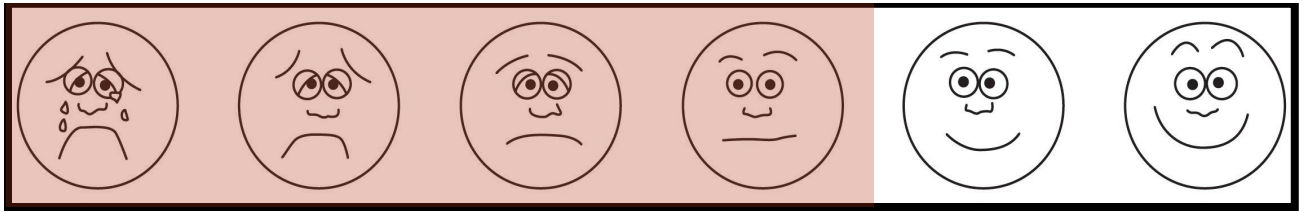


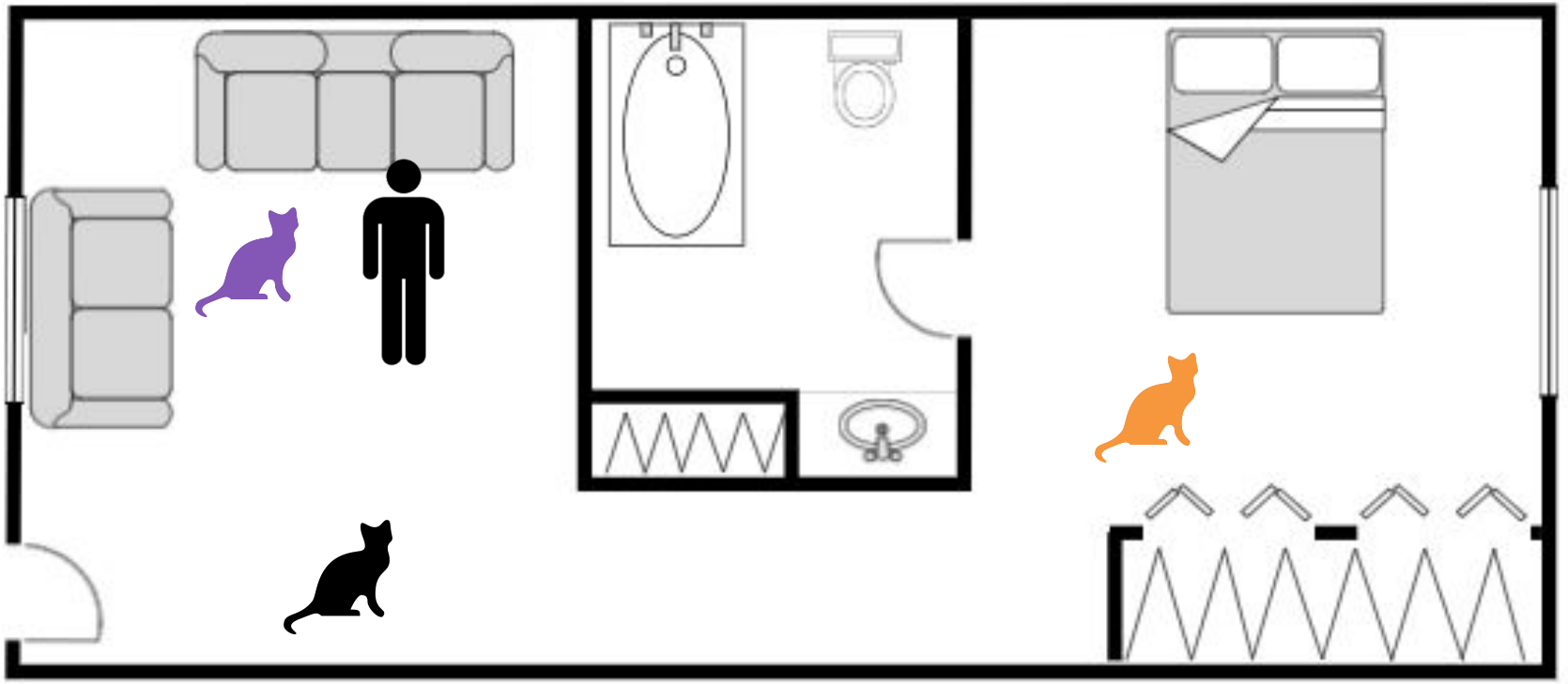


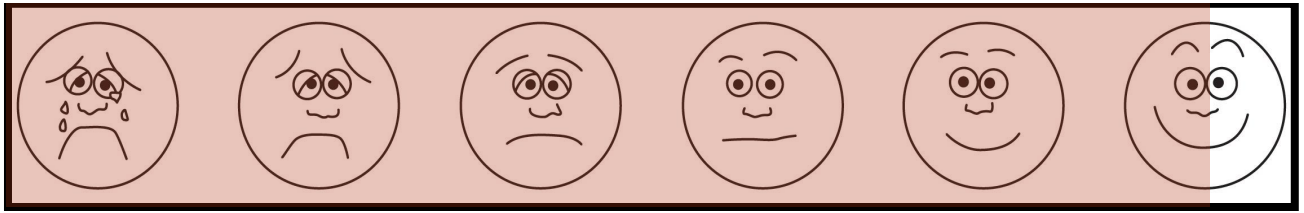












Statistics questions

- What is the relationship between each cat and happiness?
- Are those relationships statistically significant?

```
Call:
lm(formula = happiness ~  +  +  , data = dat1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8719 -1.4738  0.1606  1.3308  5.7411

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.60892    0.51750  10.839  < 2e-16 ***

            -0.16764    0.03884  -4.316 2.17e-05 ***
            -0.01025    0.06824  -0.150  0.88067
             0.17290    0.06410   2.697  0.00739 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.973 on 296 degrees of freedom
Multiple R-squared:  0.1071,    Adjusted R-squared:  0.09803
F-statistic: 11.83 on 3 and 296 DF,  p-value: 2.431e-07
```

Machine learning

- If I was to collect a new data set, could the model on the previous slide be used to predict my happiness?
  - Do the betas generalize
  - Just part of a 2-fold cross validation


Mean absolute error = 1.7
Root of mean square error = 2.15
12
8
4
0
Actual Happiness
4
5
6
7
Predicted Happiness

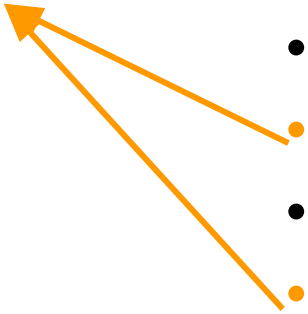
Deeper comparison

- Machine learning
  - User friendly
  - Good predictions
  - Automated
  - Black box
  - Overfitting worry
  - Generalizability
  - Lots of data
- Statistical Modeling
  - Parsimony
  - Interpretability
  - Knowledge creation
  - Inference
  - Population
  - Generalizability
  - Little data

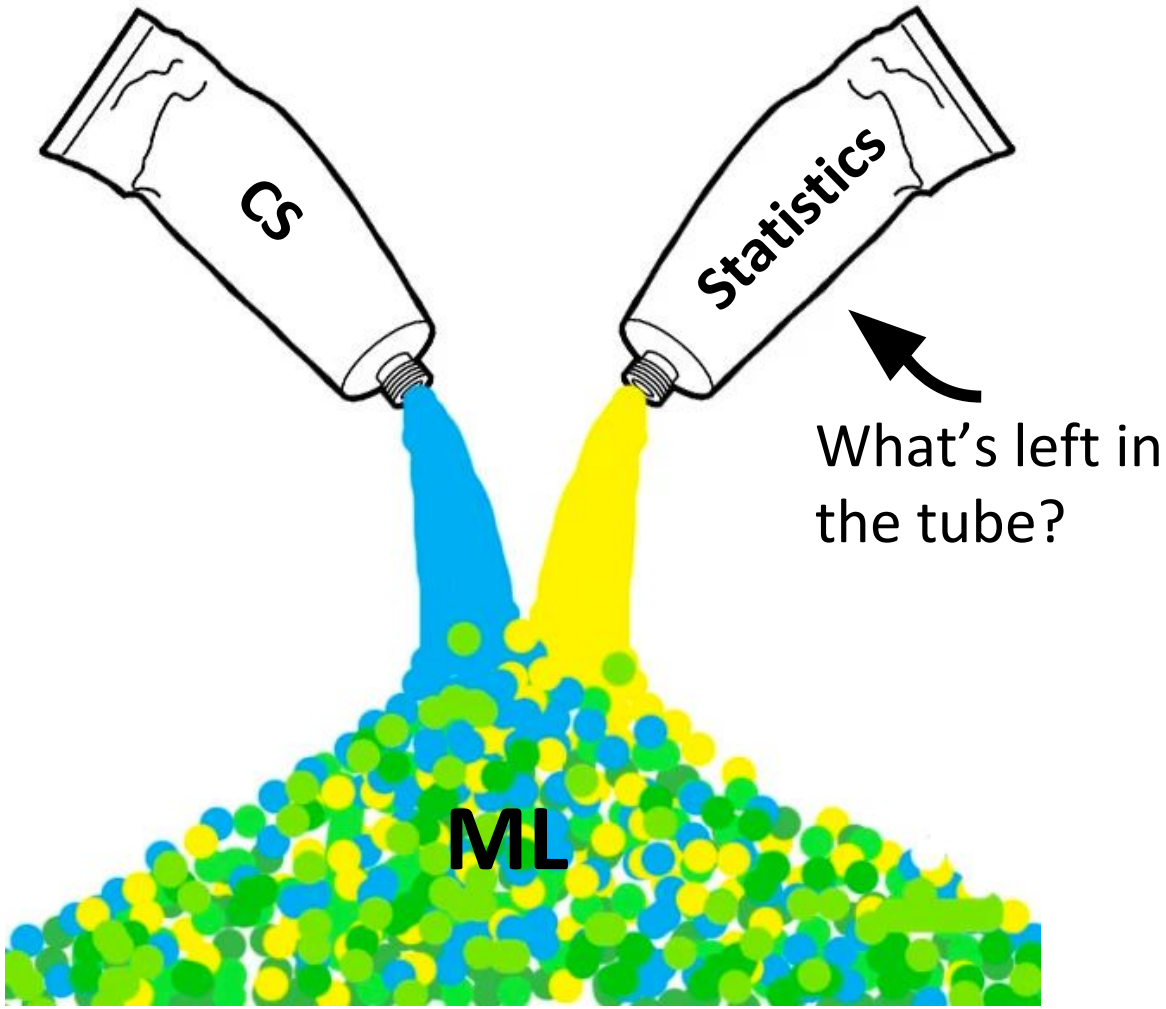
From Brian Caffo's Machine Learning vs Classical Statistics video
https://www.youtube.com/watch?v=U0XIBBuJal4&t=101s

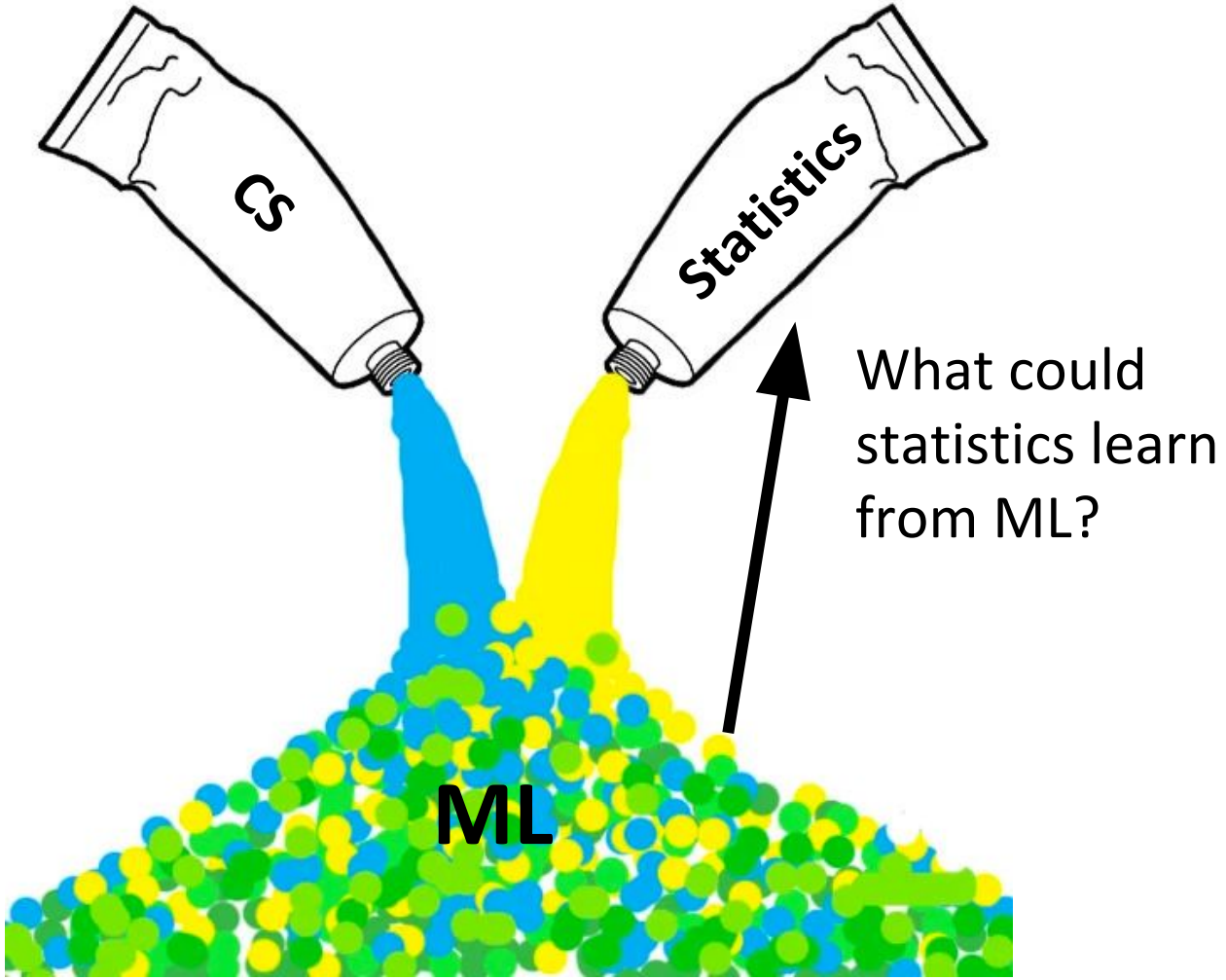
Deeper comparison

- Machine learning
  - User friendly
  - Good predictions
  - Automated
  - Black box
  - Overfitting worry
  - Generalizability
  - Lots of data
- Statistical Modeling
  - Parsimony
  - Interpretability
  - Knowledge creation
  - Inference
  - Population
  - Generalizability
  - Little data



From Brian Caffo's Machine Learning vs Classical Statistics video
https://www.youtube.com/watch?v=U0XIBBuJal4&t=101s


CS
Statistics
What's left in the tube?
ML

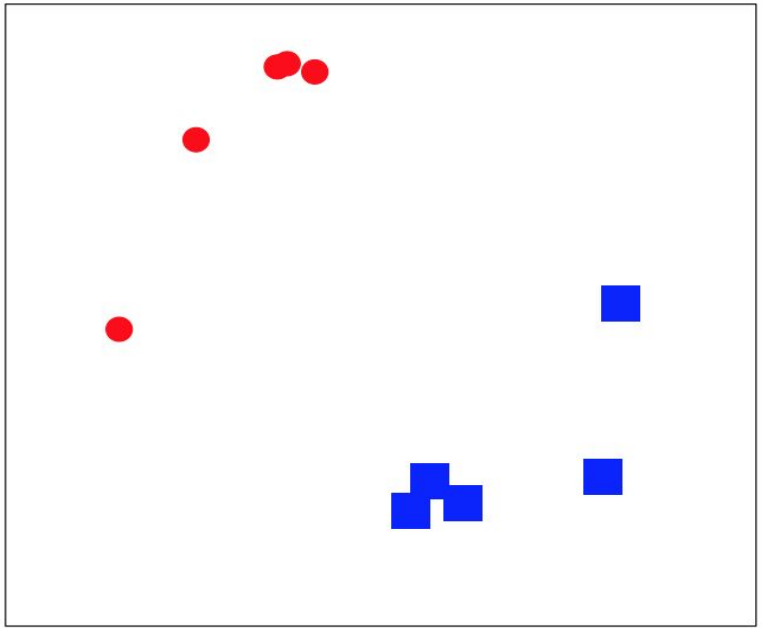

CS
Statistics
ML
What could statistics learn from ML?

Comparison

- ML
  - Papers driven by conferences
    - Faster paced
    - Creates a synergy in the field
  - Has dominated some stats topics that are currently ignored in stats
- Statistics
  - Contains important topics that could be useful in ML
    - Worry about uncertainty

Example: Semisupervised inference

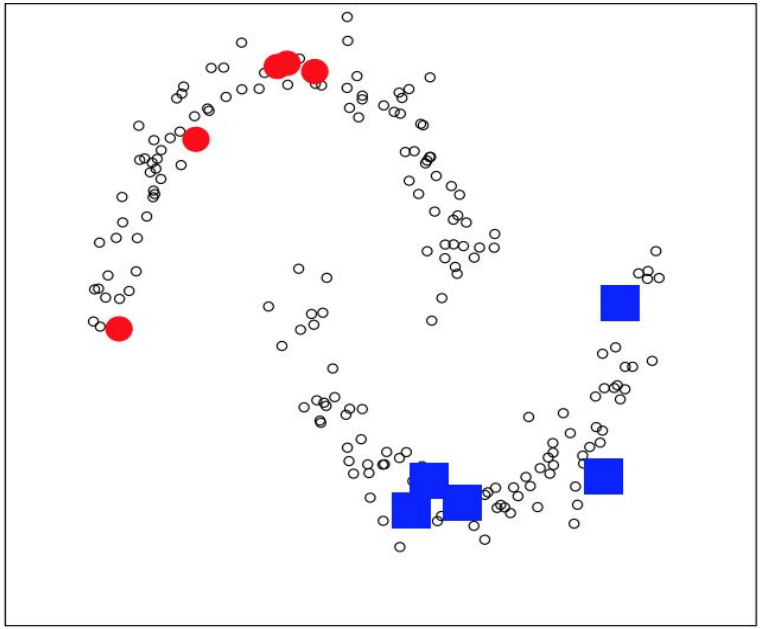
- You could try to use these data to make predictions



https://projecteuclid.org/download/pdfview_1/euclid.aos/1368018172

Example: Semisupervised inference

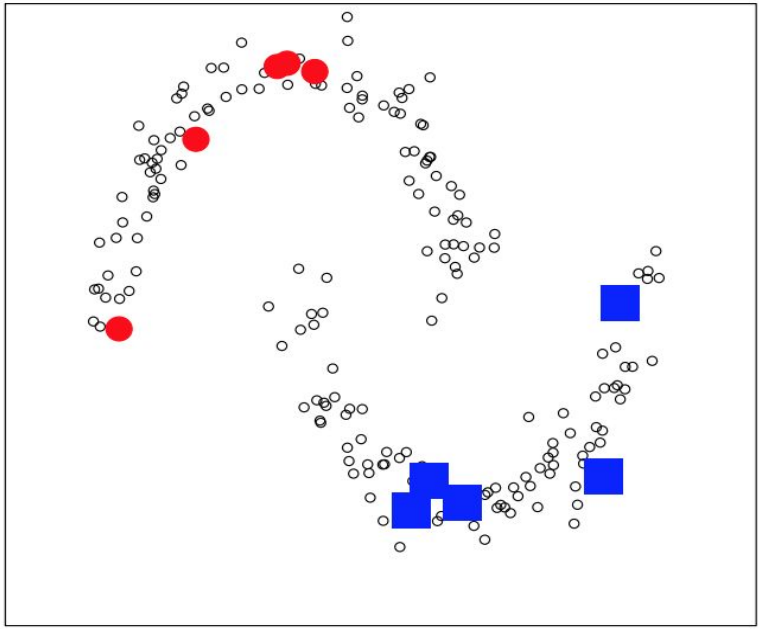
- You could try to use these data to make predictions
- Unlabeled data may help!



https://projecteuclid.org/download/pdfview_1/euclid.aos/1368018172

Example: Semisupervised inference

- You could try to use these data to make predictions
- Unlabeled data may help!
- Inference helps evaluate the assumption
  - Should unlabeled data be used?



https://projecteuclid.org/download/pdfview_1/euclid.aos/1368018172

More useful to think about how the two fields can benefit from each other

Although….

- You must consider these differences if you’re making career path choices

Main ideas

- Is machine learning just fancy statistics?
- Is machine learning better?
- How are people using ML in neuroscience?
- Things to consider when transitioning from voxelwise stats to ML

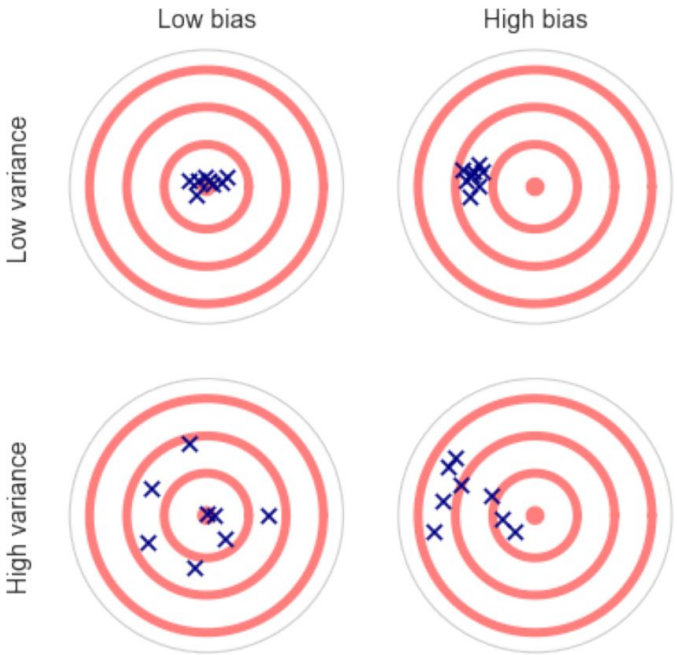
Stronger conclusions

- Asks deeper questions about model coefficients
  - Y = B0 + B1 X1 + B2X2
    - Regression focus: The model
    - ML focus: Do the betas generalize?

- Get closer to understanding the true usefulness of our theory

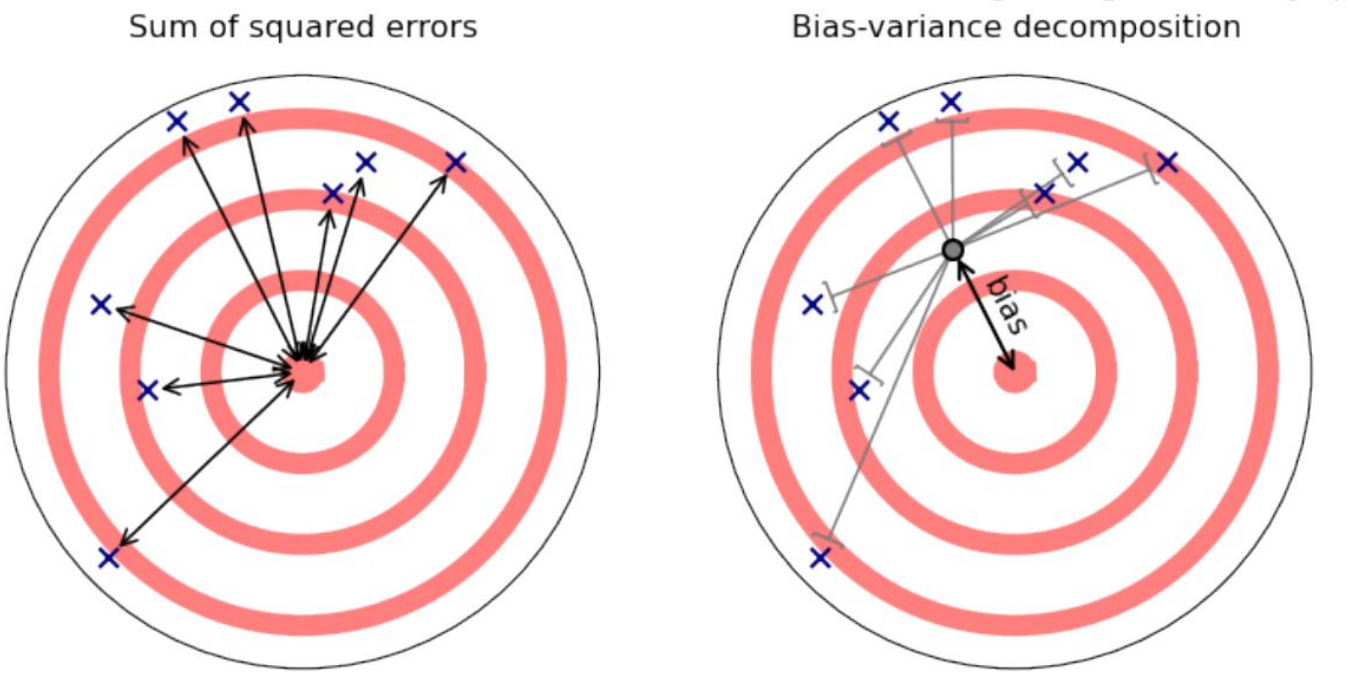
http://jakewestfall.org/publications/Yarkoni_Westfall_choosing_prediction.pdf

Bias-variance tradeoff



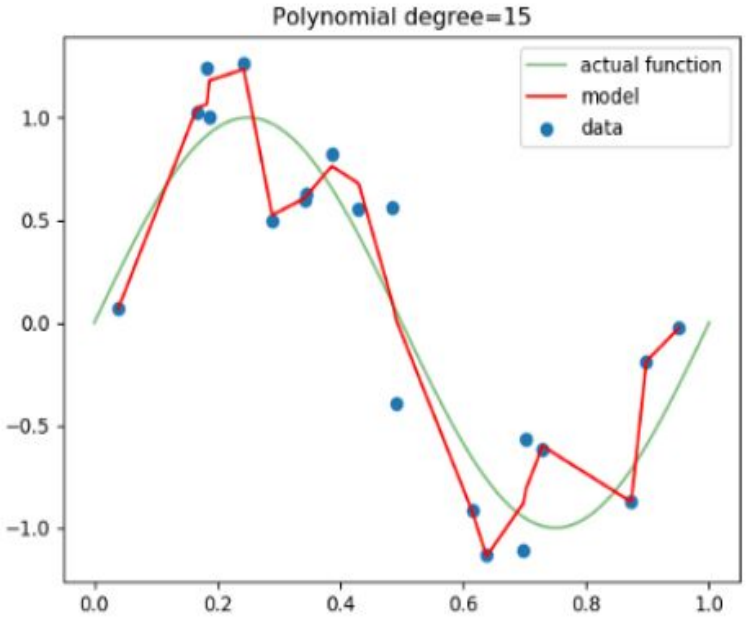
http://jakewestfall.org/publications/Yarkoni_Westfall_choosing_prediction.pdf

Bias-variance tradeoff



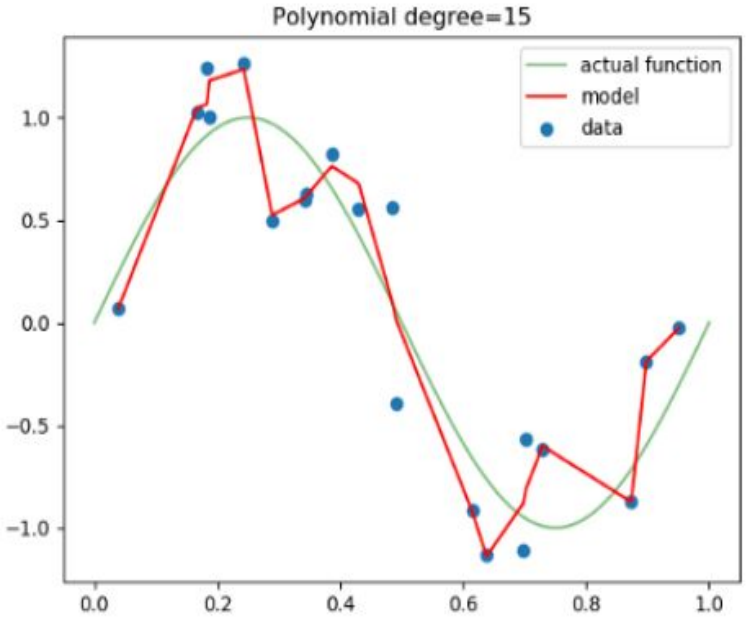
http://jakewestfall.org/publications/Yarkoni_Westfall_choosing_prediction.pdf

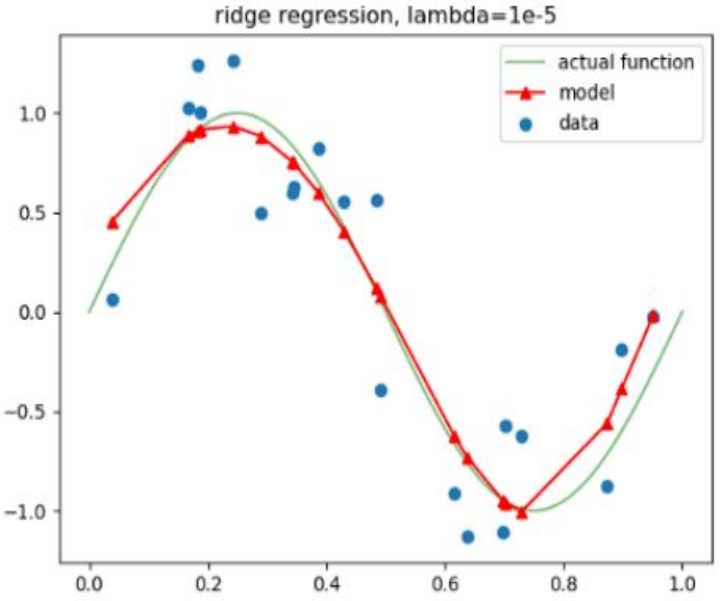
What does this look like?



Low bias and high variance regression fit
(no regularization)

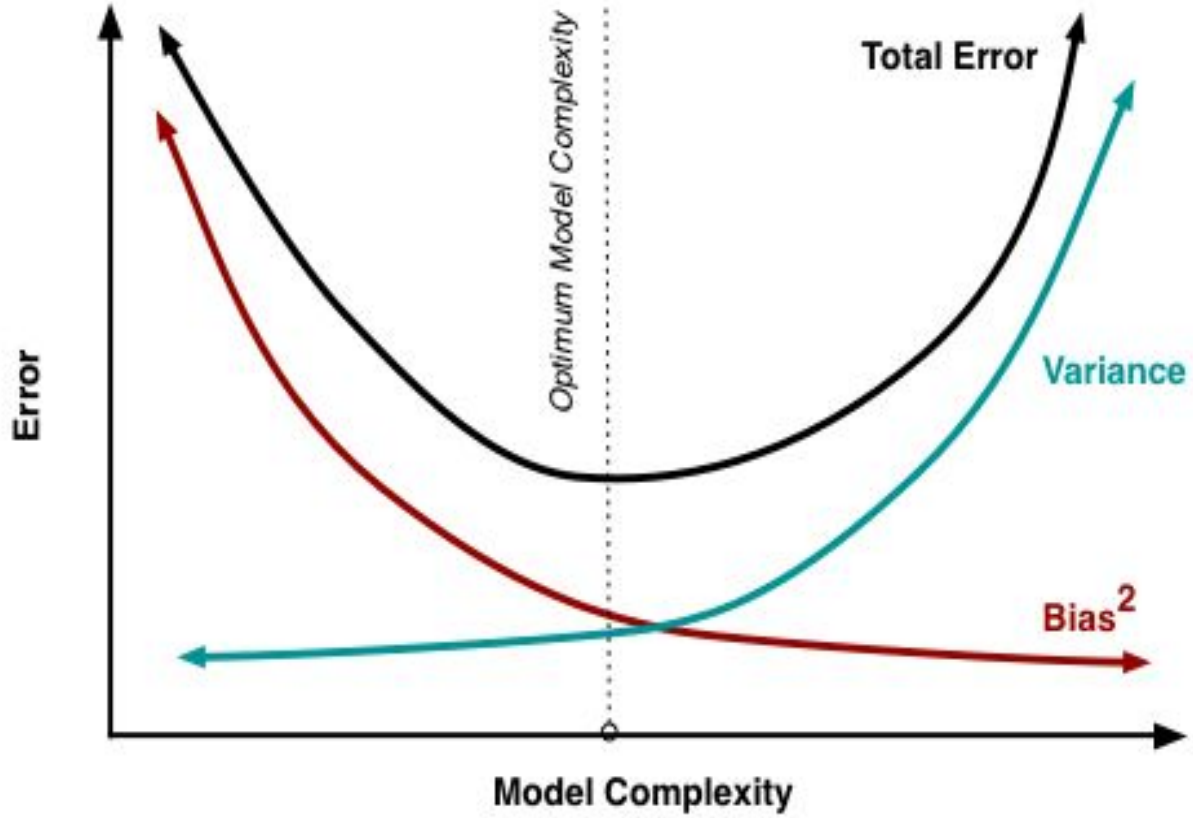
What does this look like?



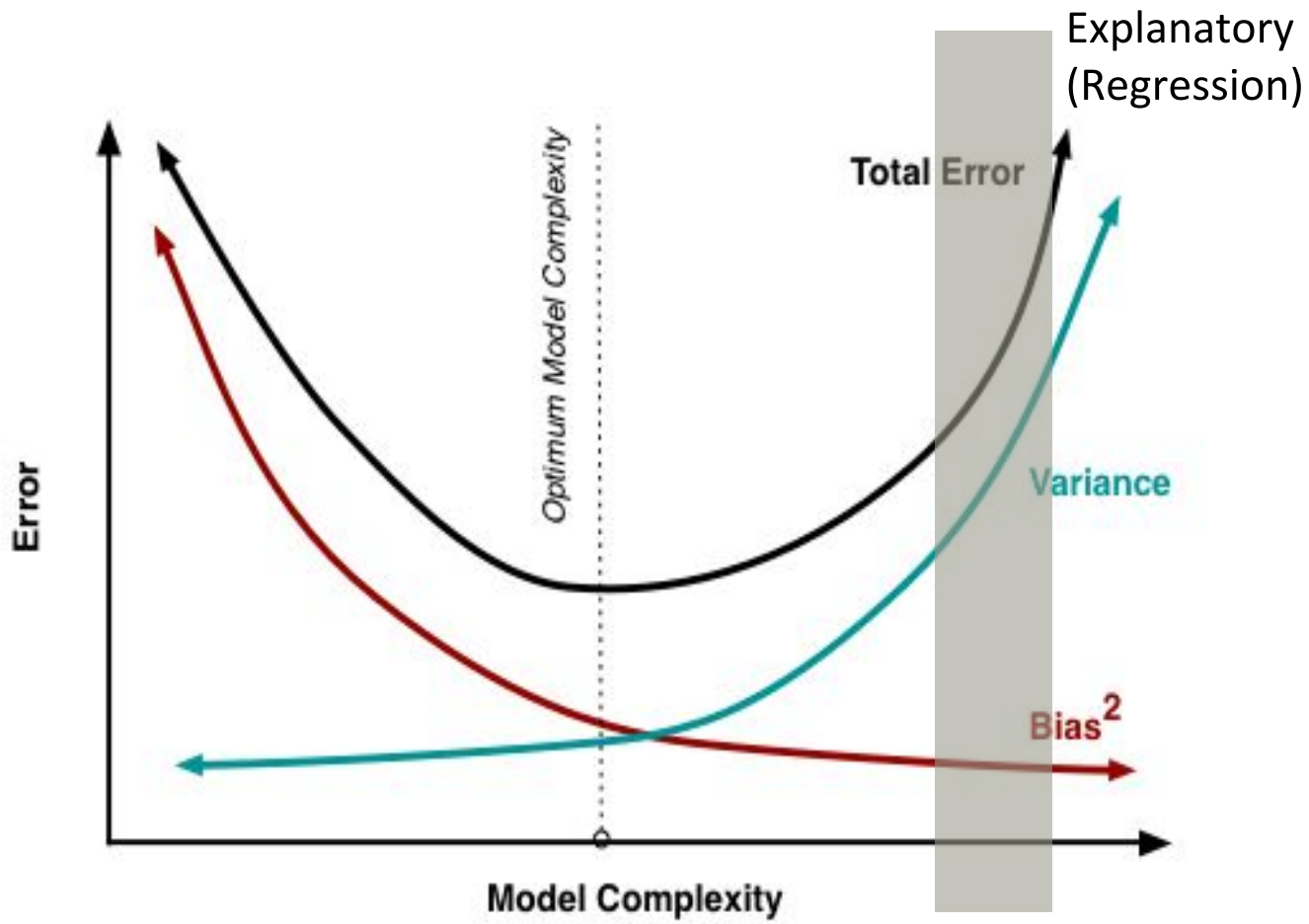


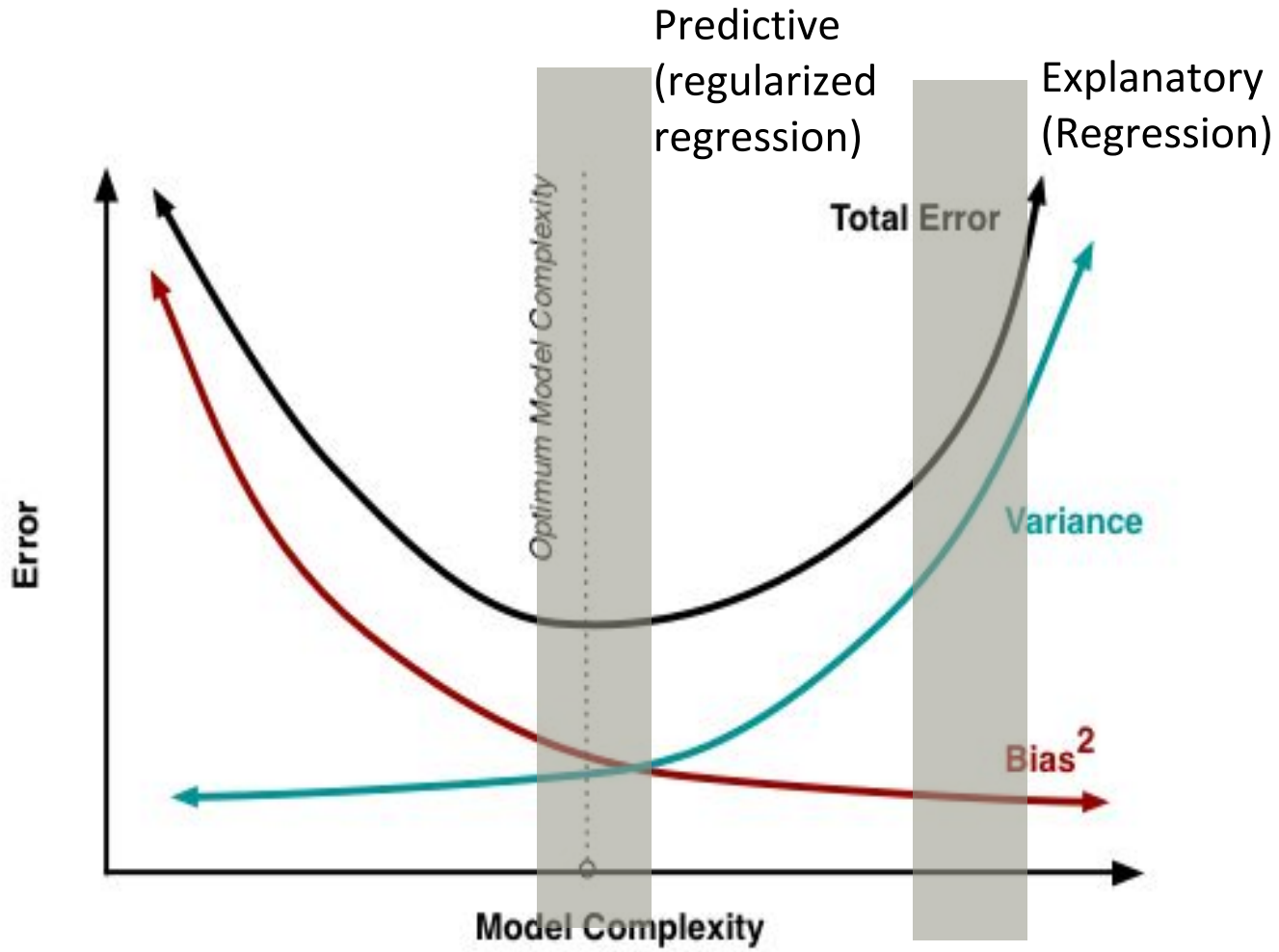
Low bias and high variance regression fit (no regularization)

Bias some betas to 0 (ridge regression) to lower the variance -> better predictions


Error
Optimum Model Complexity
Total Error
Variance
Bias^2
Model Complexity

Explanatory
(Regression)




Predictive
(regularized
regression)
Explanatory
(Regression)
Optimum Model Complexity
Total Error
Variance
Bias^2
Error
Model Complexity

Voxelwise/ROI analyses aren't reliable

Research Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
1–15

DOI: 10.1177/0956797620916786
www.psychologicalscience.org/PS
SAGE

What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis

Maxwell L. Elliott[1], Annchen R. Knodt[1], David Ireland[2], Meriwether L. Morris[1], Richie Poulton[2], Sandhya Ramrakha[2], Maria L. Sison[1], Terrie E. Moffitt[1,3,4,5], Avshalom Caspi[1,3,4,5], and Ahmad R. Hariri[1]

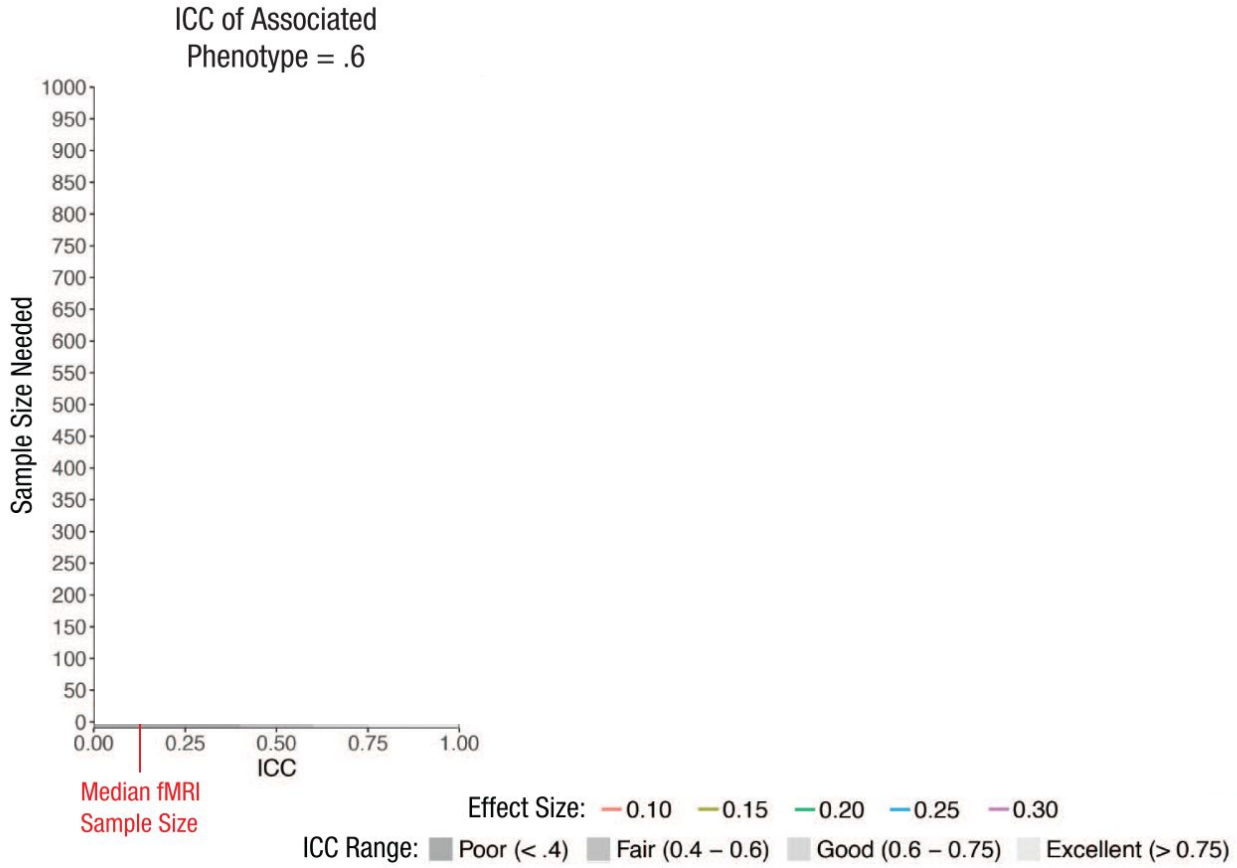


Fig. 1. The influence of task-functional MRI (fMRI) test-retest reliability on the sample size required for 80% power to detect brain–behavior correlations of effect sizes commonly found in psychological research. Power curves are shown for three levels of reliability of the associated behavioral or clinical phenotype. The figure was generated using the *pwr.r.test* function in R (Champely, 2018), with the value for *r* specified according to the attenuation formula in the Appendix. ICC = intraclass correlation coefficient.

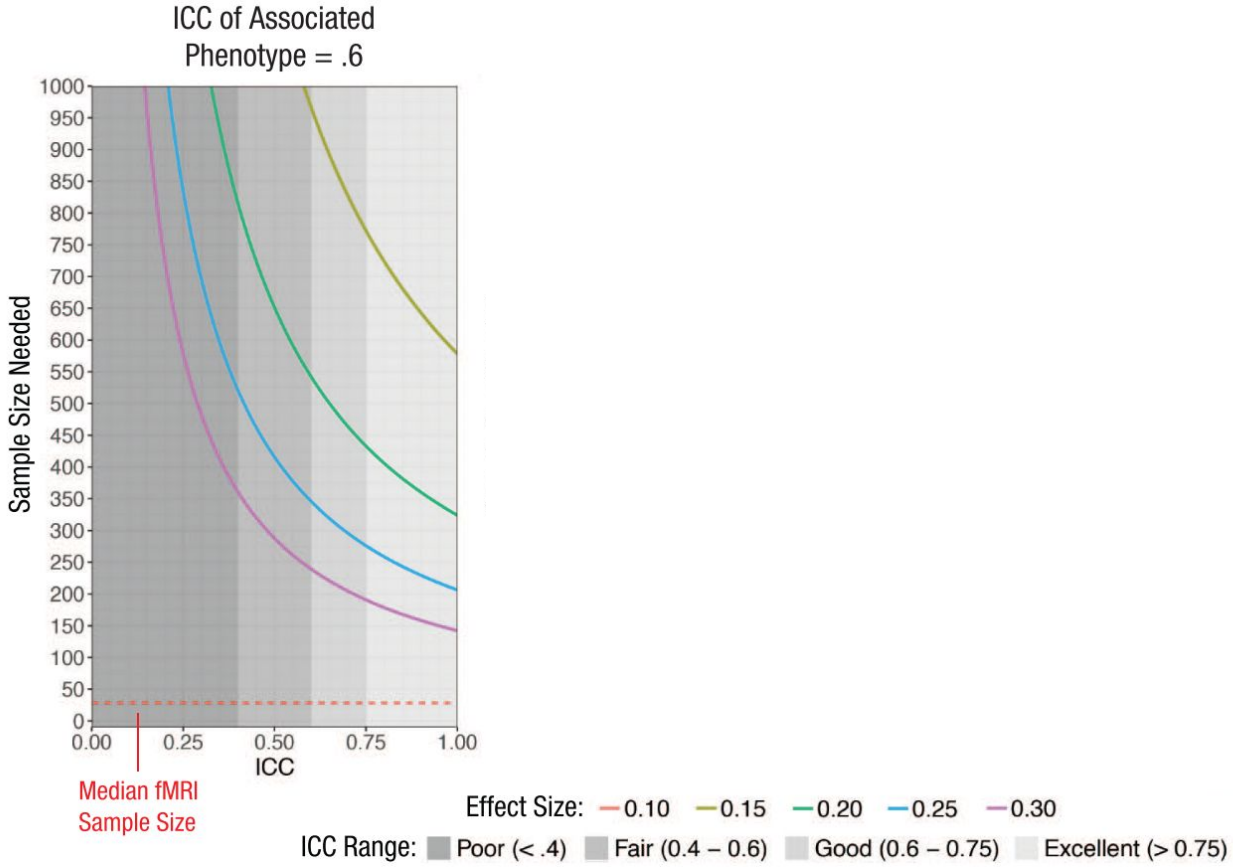


Fig. 1. The influence of task-functional MRI (fMRI) test-retest reliability on the sample size required for 80% power to detect brain–behavior correlations of effect sizes commonly found in psychological research. Power curves are shown for three levels of reliability of the associated behavioral or clinical phenotype. The figure was generated using the *pwr.r.test* function in R (Champely, 2018), with the value for *r* specified according to the attenuation formula in the Appendix. ICC = intraclass correlation coefficient.

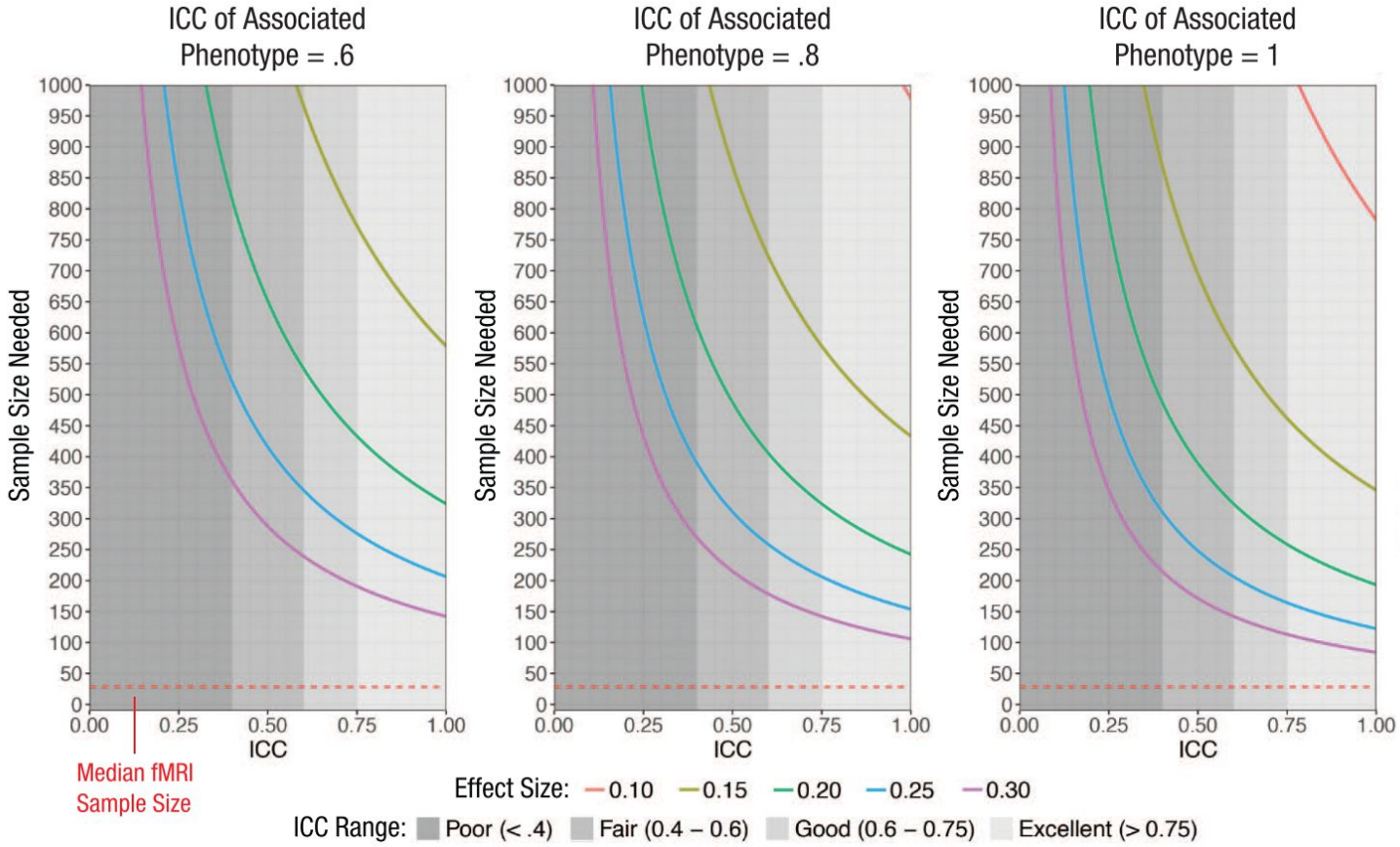


Fig. 1. The influence of task-functional MRI (fMRI) test-retest reliability on the sample size required for 80% power to detect brain–behavior correlations of effect sizes commonly found in psychological research. Power curves are shown for three levels of reliability of the associated behavioral or clinical phenotype. The figure was generated using the *pwr.r.test* function in R (Champely, 2018), with the value for *r* specified according to the attenuation formula in the Appendix. ICC = intraclass correlation coefficient.

Yet...

fMRI can be highly reliable, but it depends on what you measure

Philip A. Kragel[1*]
Xiaochun Han[2]
Thomas E. Kraynak[3]
Peter J. Gianaros[3]
Tor D. Wager[2*]

[1] Emory University
[2] Dartmouth College
[3] University of Pittsburgh

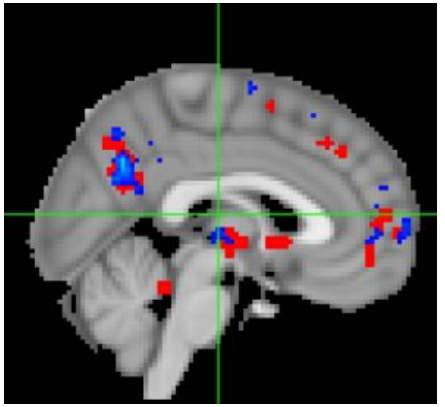
Main ideas

- Is machine learning just fancy statistics?
- Is machine learning better?
- How are people using ML in neuroscience?
- Things to consider when transitioning from voxelwise stats to ML

It isn’t quite this bad

Often cannot interpret parameters as you're used to

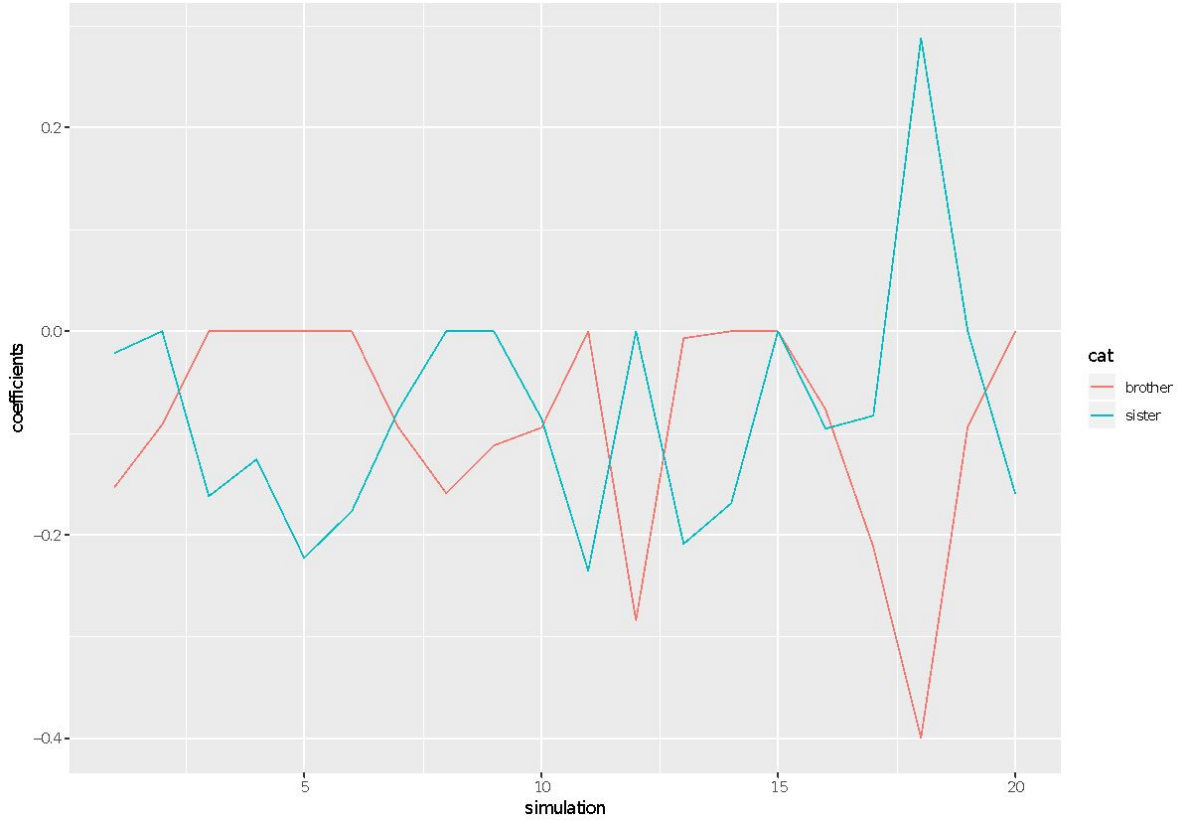
- Depends on how the regularization was done!
- The goal of the classifier is to classify
  - Doesn't necessarily matter what features it zeroed in on



Example: Lasso regression

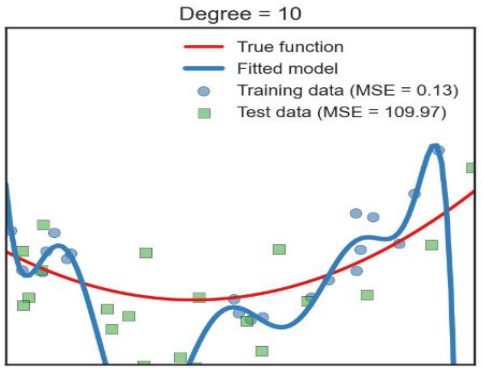
- Two cats are always right next to each other
- Their distance measures are correlated at 0.95
- Each cat’s true regression parameter is -0.1
  - For every 10 feet of increase in distance, my happiness drops by 1 :(

Lasso coefficients over multiple data sets

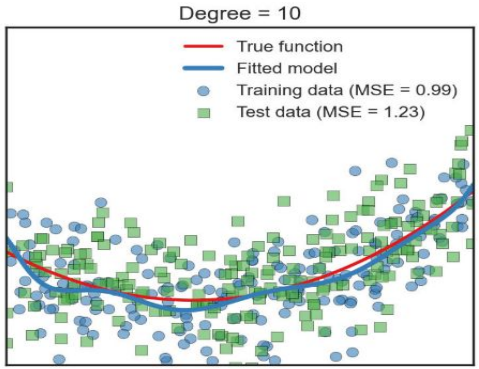


You need *more* data

- "Big" data
  - Not wide (although tend to be wider)
  - Must have more observations (longer)



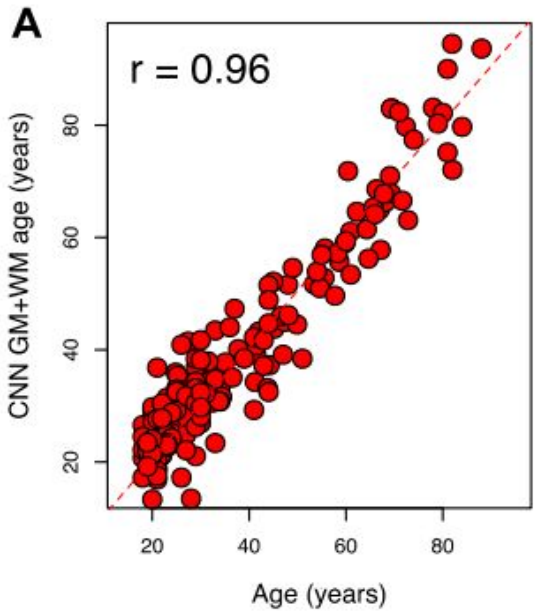
N=20



N=200

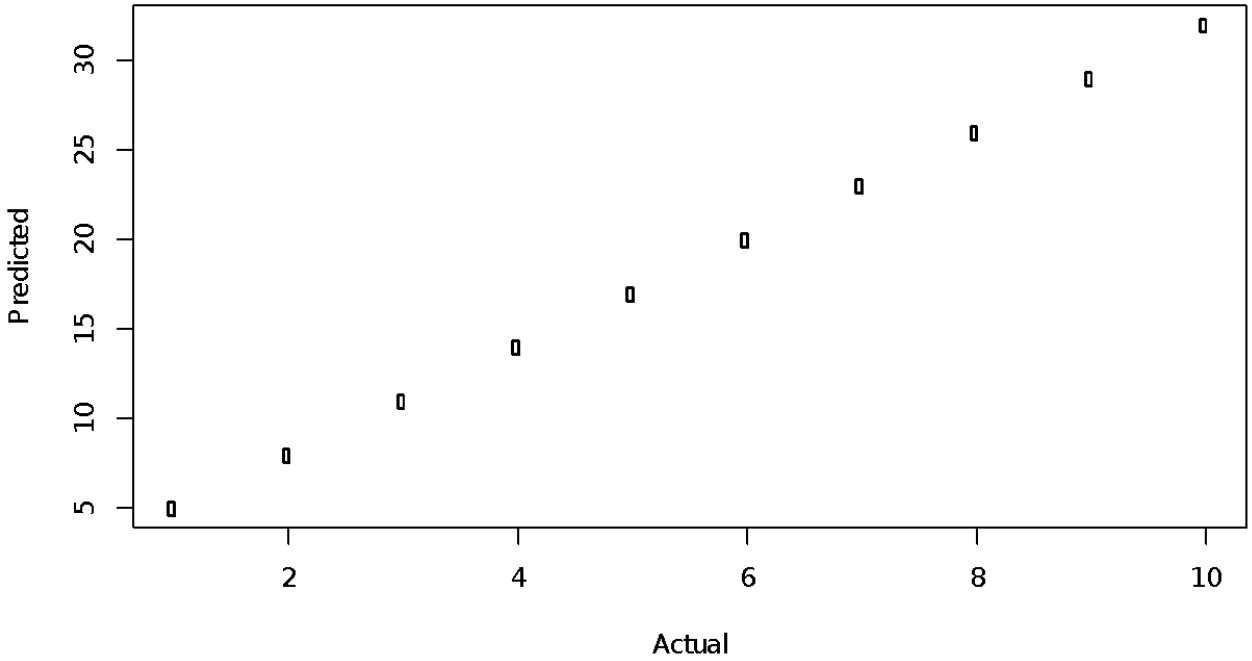
Don’t (only) use correlation to assess prediction of continuous measures

- Older folks’ predictions are *older* and younger folks’ are *younger*
- Correlation is high
  - They used plenty of other measures to assess



https://arxiv.org/abs/1612.02572

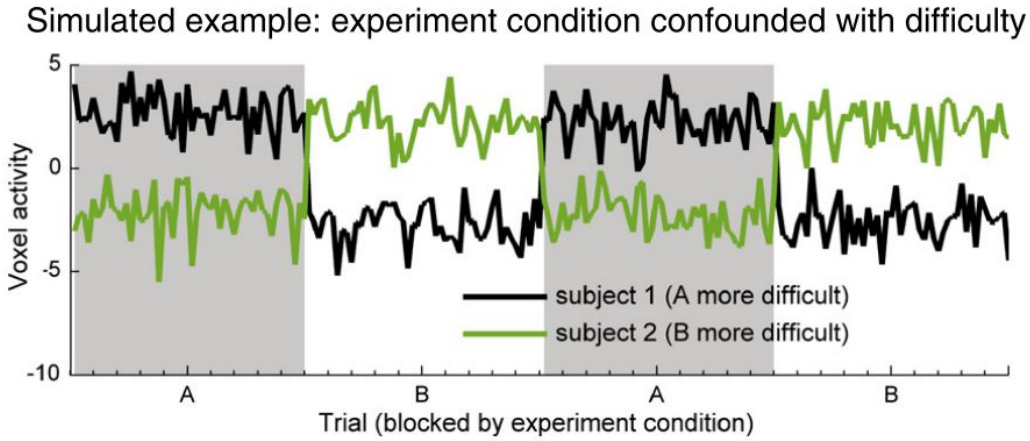
Perfect correlation, bad prediction



Cross-validation isn’t a cure-all

- Still can overfit data
- Need a left out validation set
  - You’ll only validate *once*
- If you’re developing new features, you may need a 3rd data set!

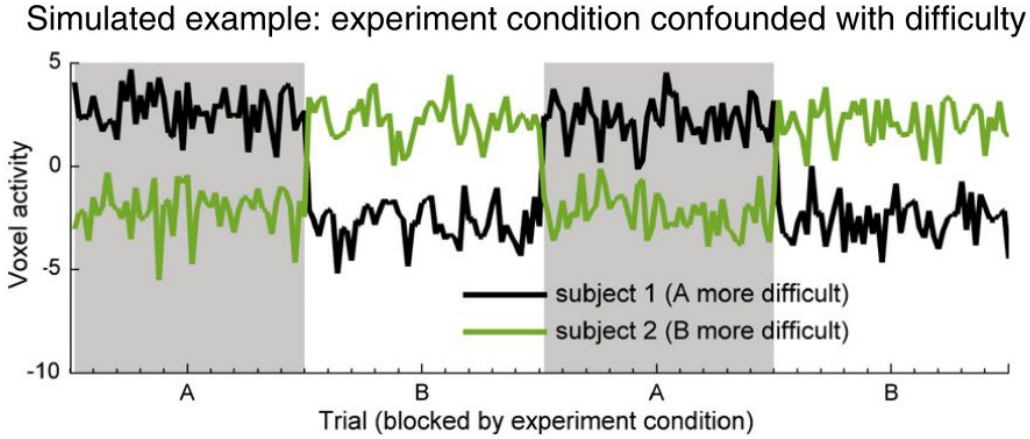
Confounds impact analyses differently



Example (I made this up): Compare math to reading, some people might find math more difficult
On average, difficulty effect should average out to reveal math/reading effect

Todd et al, 2013. https://www.sciencedirect.com/science/article/pii/S1053811913002887

Confounds impact analyses differently



Individual-Subject Summary Statistics

Subject	Experimental Effect (GLM)
Subject 1	mean(A)-mean(B) = +4.75
Subject 2	mean(A)-mean(B) = -5.56

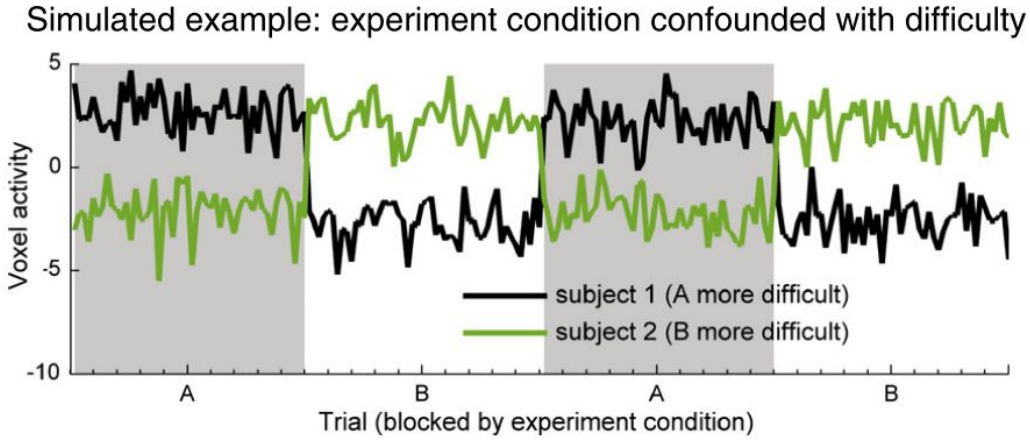
Group Test Statistics (two-tailed *t*-test)

Experimental Effect (GLM)

mean(A)-mean(B):
t_1=-0.0780, p=0.9504, n.s.

Todd et al, 2013. https://www.sciencedirect.com/science/article/pii/S1053811913002887

Confounds impact analyses differently



Individual-Subject Summary Statistics

Subject	Experimental Effect (GLM)	Discrimination Success (MVPA)
Subject 1	mean(A)-mean(B) = +4.75	classification accuracy = +13.15, within-minus-across = +3.826
Subject 2	mean(A)-mean(B) = -5.56	classification accuracy = +13.44, within-minus-across = +3.848

Group Test Statistics (two-tailed *t*-test)

Experimental Effect (GLM)	Discrimination Success (MVPA)
mean(A)-mean(B): t_1=-0.0780, p=0.9504, n.s.	classification accuracy: t_1=94.0, p<0.01, sig. within-minus-across: t_1=348, p<0.01, sig.

Todd et al, 2013. https://www.sciencedirect.com/science/article/pii/S1053811913002887

Dive in! Ask questions and be careful

Questions?